EADIndexes et IndexCorrector : mode d'emploi

Introduction

Ce petit développement part d'un constat simple : nous faisons tous des erreurs. De plus, nous ne disposons pas d'outil collaboratif (centralisé) permettant de produire de la donnée normalisée respectant un dictionnaire regroupant des termes de vocabulaire du domaine. Il existe néanmoins des solutions pour pallier ce problème : la fonction rechercher-remplacer, disponible dans tous les éditeurs de texte du marché, libres ou commerciaux.

Cependant, une question s'impose. Comment avoir une vue d'ensemble de ces informations au cours du travail ou à la fin ? Il y a deux politiques :

- on écrit bien au fur et à mesure : oXygen le permet en partie avec le mode apprentissage, mais qui ne s'applique qu'aux attributs

- on vérifie à la fin : méthode la plus répandue et utilisée (humaine)

C'est pour pallier ce second point que l'outil a été développé : il doit être utilisé conjointement avec l'outil d'extraction des index pour être parfaitement efficace. On voit donc apparaitre deux phases :

- phase 1 : obtenir la liste des index d'un document XML

- phase 2 : fournir la liste corrigée et le document XML à corriger

Cet outil est disponible en mode web, via le portail Paprika, accessible uniquement sur le réseau interne de la MSH de Dijon. En revanche, il est utilisable pour n'importe quel fonds puisqu'il la méthode s'appuie sur une organisation hiérarchique proposée par la spécification XML EAD.

Hypothèses :

On suppose que l'on dispose d'un fichier texte correctement syntaxé selon le format XML EAD. On souhaite produire la liste des mots-clés utilisés par les différentes personnes qui sont intervenues sur le fichier ou contrôler les données extraites de bases de données hétérogènes et introduites de façon automatique ou non au document XML. Les données vérifiées proviennent principalement des balises situées au sein de la balise <controlaccess />, avec:

- les noms de personnes, typiquement ce que contient la balise <persname />

- les noms d'organismes, typiquement ce que contient la balise <corpname />

- les noms géographiques, typiquement ce que contient la balise <geogname />

- les sujets, typiquement ce que contient la balise <subject />

- etc.

Données en entrée :

Un fichier XML EAD.

Données en sortie :

Un fichier Microsoft Excel, classeur constitué d'autant de feuilles qu'il y a de balises testées.

Chaque feuille du classeur est organisée à partir de six colonnes fonctionnant en trio.

Les trois premières (A, B, C), nommées respectivement WORSE_TAG, WORSE_ATTR et WORSE_VAL contiennent les données d'index issues du fichier XML, respectivement le nom du tag, le nom et la valeur de l'attribut, la valeur de la balise.

A la charge de l'utilisateur de compléter les trois colonnes suivantes (D, E, F), nommées BETTER_TAG, BETTER_ATTR, BETTER_VAL où l'on retrouve respectivement les valeurs corrigés.

Cela permet de corriger aussi bien des noms de balises ou d'attributs que leurs valeurs respectives, par lot, et ce de façon automatique. Les valeurs pouvant également être multivaluées, il est donc possible de recréer des balises par lot par ce biais. Un choix a été fait concernant les lignes non corrigées : plutôt que de ne rien faire dans ce cas, l'attribut ou le tag complet sont complètement supprimés. Il est donc plutôt conseillé de supprimer les lignes lorsque les données dans A, B et C sont correctes pour ne pas les traiter inutilement.

	A	B	C	D	E	F
1	WORSE_TAG	WORSE_ATTI	WORSE_VAL	BETTER_TAG	BETTER_ATTR	BETTER_VAL
2	corpname	role=organis	Association Republicaine des Anciens Combattants	corpname	role=organisa	Association Républicaine des Anciens Combatt
3	corpname	role=organis	Comité Executif de l'Internationale Communiste	corpname	role=organisa	Comité Executif de l'Internationale Communist
4	corpname	role=organis	Confédération Générale du Travail Unitaire	corpname	role=organisa	Confédération Générale du Travail Unitaire
5	corpname	role=organis	Internationale Communiste	corpname	role=organisa	Internationale Communiste
6	corpname	role=organis	SFIO	corpname	role=organisa	Section Française de l'Internationale Ouvrière
7	corpname	role=organis	2e IS	corpname	role=organisa	Internationale Ouvrière
8	corpname	role=organis	3e IC	corpname	role=organisa	Internationale Communiste
9	corpname	role=organis	Action Française	corpname	role=organisa	Action Française
10	corpname	role=organis	Action Socialiste	corpname	role=organisa	Action Socialiste
11	corpname	role=organis	Alliance Coopérative Internationale	corpname	role=organisa	Alliance Coopérative Internationale
12	corpname	role=organis	Amicale Populaire	corpname	role=organisa	Amicale Populaire
13	corpname	role=organis	Amis de l'URSS	corpname	role=organisa	Amis de l'URSS
14	corpname	role=organis	ARAC	corpname	role=organisa	Association Républicaine des Anciens Combatt
15	corpname	role=organis	Artisans de l'Unité	corpname	role=organisa	Artisans de l'Unité
16	corpname	role=organis	Assaciation Républicaine des Anciens Combattants	corpname	role=organisa	Association Républicaine des Anciens Combatt
17	corpname	role=organis	Association Amicale des Malgaches en France	corpname	role=organisa	Association Amicale des Malgaches en France
18	corpname	role=organis	Association des Anciens Combattants Republicains	corpname	role=organisa	Association Républicaine des Anciens Combatt
19	corpname	role=organis	Association des Anciens Combattants Républicains	corpname	role=organisa	Association Républicaine des Anciens Combatt
20	corpname	role=organis	Association des Ecrivains et Artistes Révolutionnaires	corpname	role=organisa	Association des Ecrivains et Artistes Révolutior
21	corpname	role=organis	Association des Ecrivains et Artistes Révolutionnaires	corpname	role=organisa	Association des Ecrivains et Artistes Révolutior
22	corpname	role=organis	Association des Ecrivains Revolutionnaires	corpname	role=organisa	Association des Ecrivains et Artistes Révolutior
23	corpname	role=organis	Association des Jeunesses Révolutionnaires	corpname	role=organisa	Association des Jeunesses Révolutionnaires
24	corpname	role=organis	Association des Travailleurs sans Dieu	corpname	role=organisa	Association des Travailleurs sans Dieu
25	corpname	role=organis	Association Internationale des Travailleurs	corpname	role=organisa	Association Internationale des Travailleurs
26	corpname	role=organis	Association Juridique Internationale	corpname	role=organisa	Association Juridique Internationale
27	corpname	role=organis	Association Louise Michel	corpname	role=organisa	Association Louise Michel

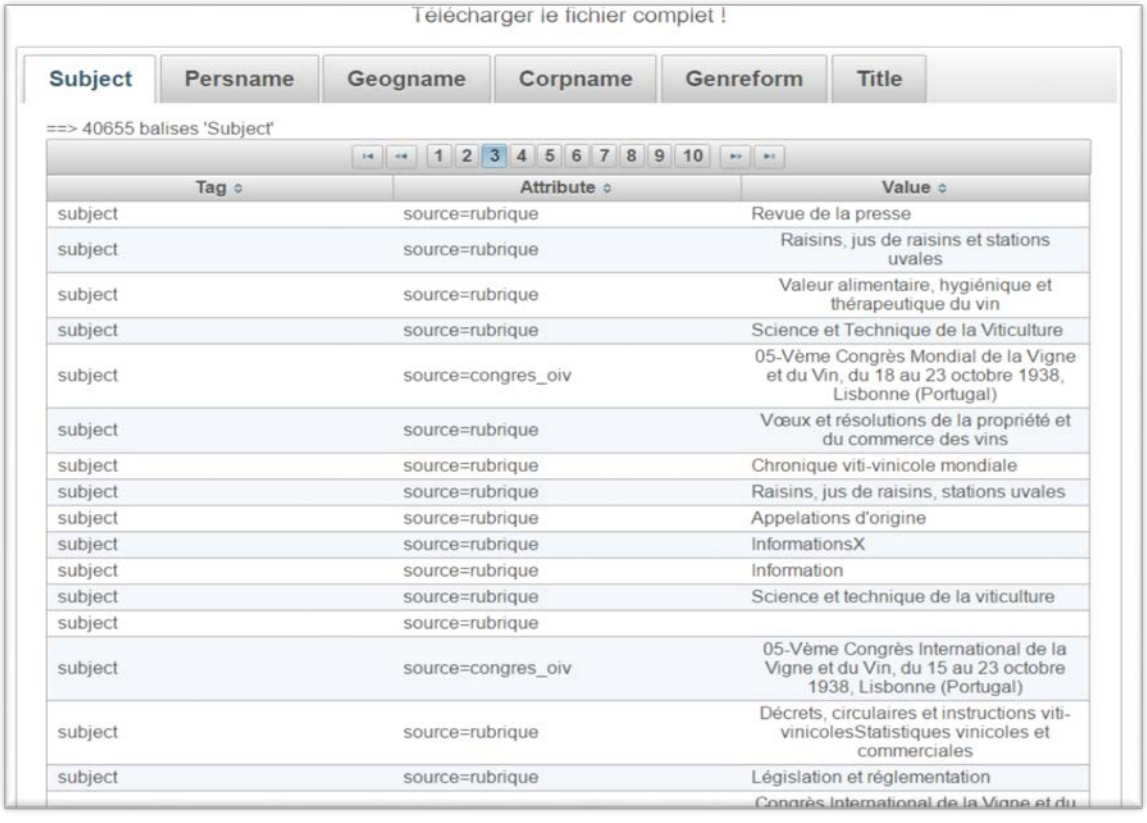
(Copie d'écran nr. 1 : index corpname du fonds « Section française de l'Internationale communiste »)

Phase 1 : extractions des attributs, valeurs d'attributs et valeurs d'index d'un fichier

En sélectionnant depuis l'interface web le fichier XML-EAD à traiter, on obtient les listes des termes et mots-clés contenus dans l'instrument de recherche, chacune des listes correspondant à un élément EAD situé au sein de la balise <controlaccess />, via le bouton « Export des index (EAD) » (voir copie d'écran nr. 2).

(Copie d'écran nr. 2 : interface web des outils développés pour le projet Paprik@2F, dont « Export des index » et « correction des index »)

Le fichier produit est téléchargeables via un lien (voir copie d'écran nr. 3).



Tag	Attribute	Value
subject	source=rubrique	Revue de la presse
subject	source=rubrique	Raisins, jus de raisins et stations uvales
subject	source=rubrique	Valeur alimentaire, hygiénique et thérapeutique du vin
subject	source=rubrique	Science et Technique de la Viticulture
subject	source=congres_oiv	05-Vème Congrès Mondial de la Vigne et du Vin, du 18 au 23 octobre 1938, Lisbonne (Portugal)
subject	source=rubrique	Vœux et résolutions de la propriété et du commerce des vins
subject	source=rubrique	Chronique viti-vinicole mondiale
subject	source=rubrique	Raisins, jus de raisins, stations uvales
subject	source=rubrique	Appelations d'origine
subject	source=rubrique	InformationsX
subject	source=rubrique	Information
subject	source=rubrique	Science et technique de la viticulture
subject	source=rubrique	
subject	source=congres_oiv	05-Vème Congrès International de la Vigne et du Vin, du 15 au 23 octobre 1938, Lisbonne (Portugal)
subject	source=rubrique	Décrets, circulaires et instructions viti-vinicolesStatistiques vinicoles et commerciales
subject	source=rubrique	Législation et réglementation
subject	source=rubrique	Congrès International de la Vigne et du

(Copie d'écran nr. 3 : téléchargement du tableur produit à partir du fichier XML-EAD)

Phase 2: réalisation des corrections proprement dites

Pour voir la marche à suivre détaillé du processus de correction, voir ci-après « Etude de cas : correction de l'élément « geogname » (dans le cadre du projet Bulletin de l'OIV) »

Phase 3 : import des corrections dans le fichier EAD

Dès que le fichier produit à l'étape précédente est modifié tout ou en partie seulement, il est possible de le soumettre au second outil (via le bouton « Correction des index (EAD) », (voir copie d'écran nr. 4)) afin qu'il réalise les opérations de correction souhaitées. Encore une fois, ces deux phases peuvent être répétées plusieurs fois de suite sur le même document ou des documents différents, puisque ces derniers obéissent aux mêmes règles syntaxiques.

(Copie d'écran nr. 4 : import du tableur corrigé pour mise à jour du fichier XML-EAD)

Hypothèses :

On suppose que l'on dispose du tableur composé des *n* feuilles avec les corrections apportées.

Données en entrée :

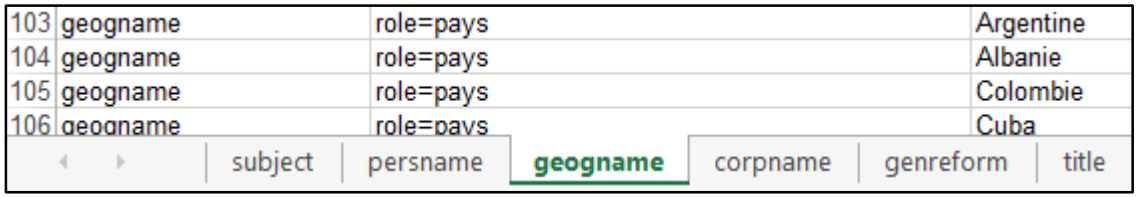
Un fichier au format Microsoft Excel corrigé.

Données en sortie :

Un fichier XML EAD mis à jour avec les corrections apportées. Théoriquement, avec cette méthode, tous les fichiers XML relevant d'une même thématique peuvent intégrer et exploiter le même vocabulaire, au moins sur les termes clés.

Etude de cas : correction de l'élément « geogname » (dans le cadre du projet Bulletin de l'OIV)

Rappel : Chaque élément présent dans le fichier XML-EAD concerné fait l'objet d'une feuille distincte (voir illustration nr. 1)



(Illustration nr. 1 : tableur généré à partir du bouton « Export des index (EAD »))

Chacune des feuilles est composée de six colonnes, l'en-tête des colonnes se déclinant de la façon suivante. Par exemple, pour la feuille geogname :

WORSE_TAG	WORSE_ATTR	WORSE_VAL	BETTER_TAG	BETTER_ATTR	BETTER_VAL

Employés pour :

WORSE_TAG : Nom d'élément (erroné ou non)

WORSE_ATTR : Nom d'attribut (erroné ou non)

WORSE_VAL : Valeur de départ (erronée ou non)

BETTER_TAG : Nom d'élément à retenir

BETTER_ATTR : Nom d'attribut à retenir

BETTER_VAL : Valeur à retenir

A noter, les informations contenues dans la ligne d'en-tête du fichier sont données à titre indicatif, pour le bon fonctionnement du programme, cette ligne peut soit, comporter ces informations, soit être laissée vide.

WORSE_TAG	WORSE_ATTR	WORSE_VAL	BETTER_TAG	BETTER_ATTR	BETTER_VAL
geogname	role=pays	Maroc			
geogname	role=pays	Mroc			

(Illustration nr. 2 : tableur avec en-tête renseignée)

geogname	role=pays	Maroc			
geogname	role=pays	Mroc			

(Illustration nr. 3 : tableur avec en-tête vide)

Chaque ligne correspond à un triplet **nom d'élément**, **nom d'attribut** et **sa valeur**, **valeur d'élément**

nom d'élément	nom d'attribut= valeur	valeur d'élément
geogname	role=ville	Madrid (Espagne)
geogname	role=pays	Finlande
geogname	role=departement	Gironde (France)
geogname	role=ville	Cognac (Charente, France)
geogname	role=etat	Massachusetts (Etats-Unis)
geogname	source=denomination	Saint-Emilion (France)

(Illustration nr. 4 : exemple de triplet)

Plusieurs types de corrections sont possibles

Corrections de valeurs

WORSE_TAG	WORSE_ATTR	WORSE_VAL	BETTER_TAG	BETTER_ATTR	BETTER_VAL
geogname	role=pays	Maroc			
geogname	role=pays	Mroc			
geogname	role=pays	maroc			

(Illustration nr. 5 : tableur généré à partir du bouton « Export des index »)

Note : au moment de l'export du fichier EAD, toutes les informations sont exportées dans le tableur, qu'elles soient correctes ou non.

WORSE_TAG	WORSE_ATTR	WORSE_VAL	BETTER_TAG	BETTER_ATTR	BETTER_VAL
geogname	role=pays	Maroc			
geogname	role=pays	Mroc	geogname	role=pays	Maroc
geogname	role=pays	maroc	geogname	role=pays	Maroc

(Illustration nr. 6 : tableur corrigé)

Marche à suivre : on renseigne dans les colonnes de droite les informations correctes, lorsque des modifications doivent être apportées.

La totalité des trois colonnes doit être renseignée.

Il est recommandé de procéder par copier/coller d'informations correctes pour ne pas provoquer de nouvelles erreurs au moment des modifications.

Tableau 2b pour import dans le fichier EAD d'origine.

Au moment de l'import sont prises en compte les valeurs des trois colonnes de droite. Il importe donc de supprimer les lignes contenant des informations valides à gauche (et rien à droite) dans le tableur. Dans le cas contraire, les informations valides au moment de l'import du tableur, seraient supprimées dans le fichier EAD d'origine.

Exemple de tableur bien renseigné :

WORSE_TAG	WORSE_ATTR	WORSE_VAL	BETTER_TAG	BETTER_ATTR	BETTER_VAL
geogname	role=pays	Mroc	geogname	role=pays	Maroc
geogname	role=pays	maroc	geogname	role=pays	Maroc

(Illustration nr. 7 : tableur pour import via le bouton « Correction des index (EAD) » pour mise à jour du fichier EAD d'origine)

Attention 

Il ne faut pas retirer les informations des lignes mais bien supprimer les lignes correctes, car si le programme rencontre une ligne vide il cesse de fonctionner.

Ainsi, ce tableau n'est pas correctement formé

WORSE_TAG	WORSE_ATTR	WORSE_VAL	BETTER_TAG	BETTER_ATTR	BETTER_VAL
geogname	role=pays	Mroc	geogname	role=pays	Maroc
geogname	role=pays	maroc	geogname	role=pays	Maroc

(Illustration nr. 8 : tableur avec une ligne vide susceptible de bloquer le logiciel)

Corrections d'attributs

Tableau 1

WORSE_TAG	WORSE_ATTR	WORSE_VAL	BETTER_TAG	BETTER_ATTR	BETTER_VAL
geogname	role=ville	Maroc			

Tableau 2

WORSE_TAG	WORSE_ATTR	WORSE_VAL	BETTER_TAG	BETTER_ATTR	BETTER_VAL
geogname	role=ville	Maroc	geogname	role=pays	Maroc

Tableau 3

WORSE_TAG	WORSE_ATTR	WORSE_VAL	BETTER_TAG	BETTER_ATTR	BETTER_VAL
geogname	role=pays	Maroc			

Corrections d'éléments

Tableau 1

WORSE_TAG	WORSE_ATTR	WORSE_VAL	BETTER_TAG	BETTER_ATTR	BETTER_VAL
persname	role=auteur	Maroc			

Tableau 2

WORSE_TAG	WORSE_ATTR	WORSE_VAL	BETTER_TAG	BETTER_ATTR	BETTER_VAL
persname	role=auteur	Maroc	geogname	role=pays	Maroc

A noter, dans cet exemple, des corrections ont été apportées dans deux colonnes distinctes

Tableau 3

WORSE_TAG	WORSE_ATTR	WORSE_VAL	BETTER_TAG	BETTER_ATTR	BETTER_VAL
geogname	role=pays	Maroc			

Phase 4 : renouvellement de l'export pour vérification

Une fois le fichier EAD corrigé, il est important de renouveler l'export vers le tableur pour vérifier qu'il ne reste plus d'erreurs.

WORSE_TAG	WORSE_ATTR	WORSE_VAL	BETTER_TAG	BETTER_ATTR	BETTER_VAL
geogname	role=ville	Madrid (Espagne)			
geogname	role=pays	Maroc			
geogname	role=departement	Gironde (France)			
geogname	role=ville	Cognac (Charente, France)			
geogname	role=etat	Massachusetts (Etats-Unis)			
geogname	source=denomination	Saint-Emilion (France)			

(Illustration nr. 9 : tableur exporté depuis le fichier EAD corrigé pour une nouvelle relecture)

Les erreurs de saisie ayant été corrigées, le fichier ne comporte plus qu'une seule ligne pour un triplet donné. Il n'y a pas de nouvelles modifications à apporter dans ce cas.

Intérêt d'un tel outil

- Le tableur permet de repérer des erreurs qui ne rendent pas le fichier XML invalide et donc n'empêchent pas sa mise en ligne mais nuisent à la qualité des métadonnées et des recherches sur le portail de publication
  Ainsi une recherche relative au Maroc ne renverra qu'un résultat avant correction au lieu de 3 après correction.
- Il est certain qu'un "rechercher/remplacer" rendrait un service équivalent, mais l'inventaire automatique des termes et la correction "en masse" facilitent considérablement la lecture et la modification.
- Il arrive qu'un instrument de recherche soit trop volumineux pour qu'il soit possible d'effectuer des manipulations de type rechercher/remplacer dans un éditeur de type oXygen